

Re: why does leave one out cross validation have high variance

Source: <http://sci.tech-archive.net/Archive/sci.math.num-analysis/2007-11/msg00078.html>

- *From:* DaleGriffiths <dsmackin@xxxxxxxxxx>
 - *Date:* Tue, 06 Nov 2007 11:48:32 -0800
-

On Oct 27, 11:56 am, Shrihari <shrihari.vasude...@xxxxxxxxxx> wrote:

Hi all,

I have a query regarding cross validation methods for estimating the generalization error of a classifier.

Most prior literature says that leave-one-out CV (LOOCV) has low bias and high variance. Low bias I completely understand. The thing I do not understand is why it should have high variance – the only explanation I have been able to gather thus far is that since the models computed in each stage of the CV are very similar to each other (obviously, because they only differ by 2 training data instances), the variance is high (this is not obvious and in fact counter intuitive as I would expect similar models to produce similar results).

Could someone please explain this to me. Thanks for any help.

Regards
Shrihari

Think of this way. If you use 10 fold cross validation, you have 10 estimates of your statistic. We can call them X_1, X_2, \dots, X_{10} . Each of these is itself an average. So you would expect the 10 measurement to be normally distributed around the expected value $\langle X \rangle$.

you would expect that values X_1 through X_{10} to be similar and approximately normally distributed around the mean $\langle X \rangle$. But in LOO CV, you have X_1, X_2, \dots, X_N where N is the number of data points in your training sample. Each X_i is not an average but single measurement. So the variance between each are much greater. The great variance is often not a problem because what you really care about is the estimate $\langle X \rangle$. LOO CV usually gives the most accurate estimate of $\langle X \rangle$, i.e. it is the least biased.

Re: why does leave one out cross validation have high variance